AD-A040 240    NORTH CAROLINA UNIV AT CHAPEL HILL DEPT OF STATISTICS    F/G 12/1
COMPLETENESS COMPARISONS AMONG SEQUENCES OF SAMPLES. II. CENSOR--ETC(U)
FEB 77   N L JOHNSON                                    DAAG29-74-C-0030
UNCLASSIFIED    MIMEO SER-1103                ARO-11959.12-M           NL

OF
ADA
040240

END
DATE
FILMED
6-77

ARO-11959.12-On

COMPLETENESS COMPARISONS AMONG SEQUENCES OF SAMPLES
II. CENSORING FROM ABOVE OR BELOW, AND GENERAL CENSORING

(12)

P.S.

TECHNICAL REPORT

N. L. JOHNSON

FEBRUARY, 1977

U. S. ARMY RESEARCH OFFICE
RESEARCH TRIANGLE PARK, NORTH CAROLINA

DAAG29-74-C-0030

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27514

DDC
JUN 6 1977
C

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS<br>BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br><br>Completeness Comparisons Among Sequences of Samples. II. Censoring from Above or Below, and General Censoring | | 5. TYPE OF REPORT & PERIOD COVERED<br><br>TECHNICAL rept. |
| | | 6. PERFORMING ORG. REPORT NUMBER<br>Mimeo Series No. 1103 |
| 7. AUTHOR(s)<br><br>N. L. Johnson | | 8. CONTRACT OR GRANT NUMBER(s)<br><br>DAH G29-74-C-0030 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Department of Statistics<br>University of North Carolina<br>Chapel Hill, North Carolina 27514 | | 10. PROGRAM ELEMENT, PROJECT, TASK<br>AREA & WORK UNIT NUMBERS |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Army Research Office<br>Box 1221<br>Research Triangle Park, North Carolina 27709 | | 12. REPORT DATE<br>February, 1977 |
| | | 13. NUMBER OF PAGES<br>22  24 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)<br>ARO    11757.12-M | | 15. SECURITY CLASS. (of this report)<br><br>UNCLASSIFIED |
| | | 15a. DECLASSIFICATION/DOWNGRADING<br>SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release: Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Censoring. Sequences of samples. Nonparametric statistics. Indicator variables.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Two sequences of samples are available; and it is known that just one sequence represents the remainder of original random samples which have been censored. Methods of deciding which sequence has been censored are described, for the cases of (a) censoring from above or below (b) general censoring. It is not necessary to know the population distribution of the observed characteristic to apply the method. There is some calculation and discussion of the accuracy of the methods.

DD FORM 1473  EDITION OF 1 NOV 68 IS OBSOLETE
1 JAN 73

182854

Completeness Comparisons Among Sequences of Samples
II. Censoring from Above or Below, and General Censoring*

by

N. L. Johnson

Department of Statistics
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27514

## 1. Introduction

As in [1] we will consider situations in which we have two sequences
of samples $A_1 \equiv A_{11}, A_{12}, \ldots$; $A_2 \equiv A_{21}, A_{22}, \ldots$ respectively. We denote
the order statistics in $A_{ij}$ by

$$X_{ij1} \leq X_{ij2} \leq \ldots \leq X_{ijr_{ij}}$$

( $r_{ij}$ is the number of observed values in $A_{ij}$ ).

For convenience we will also use the notation

$$X_{i1} \leq X_{i2} \leq \ldots \leq X_{ir_i}$$

for order statistics in a typical member of $A_i$ (i = 1,2).

We further suppose that it is known that each sample in one of the two
sequences (it is not specified which one) is a complete random sample while
the other is a random sample which has been censored by removal of some
extreme observations. We will denote the *original* sample or sequence by
placing a bar over the corresponding symbol – thus $\overline{A}_{11}$ would belong to $\overline{A}_1$ .
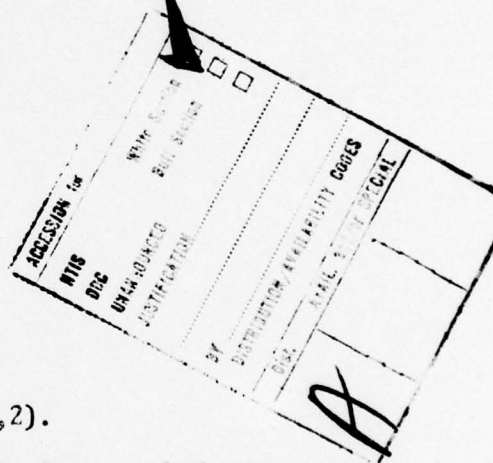
In [1] it was supposed that symmetrical censoring (removal of the s greatest and s least values from an original complete random sample of size (r+2s) ) had been used on one sequence. In the present paper we will consider censoring from above or below -- for example, removal of the greatest s *or* the least s (but not both) from an original complete random sample of size (r+s).

We will also consider a natural extension of a general purpose test of extreme sample censoring (suggested in [2] for use when the proportions of greatest and least values removed are unknown).

## 2. Censoring from Above or Below

For definiteness we will discuss censoring from above, in which the s greatest values are removed from a complete random sample of size (r+s) -- or more generally $(r_{ij} + s)$.

The cumulative distribution function, (cdf) $F_i(x)$, will be supposed to be the same for each of the i-th pair of samples $A_{i1}$ and $A_{i2}$ . It will be assumed to be absolutely continuous, with probability density function $f_i(x) = dF_i/dx$.

We will also use $F(x)$ for the cdf in a general pair of samples for $A_1$ and $A_2$ for convenience.

## 2.1 Population Distribution(s) Known

In this case the ratio for the hypotheses $H_1$ vs. $H_2$ where

$$H_i \equiv \text{sequence } A_i \text{ is censored}$$

is

$$L_1/L_2 = \prod_{j=1}^{m} \left[ -\frac{(r_{1j}+1)^{[s]}}{(r_{2j}+1)^{[s]}} \cdot \left\{ \frac{1-F(X_{1j},r_{1j})}{1-F(X_{2j},r_{2j})} \right\}^2 \right] \tag{1}$$

(remember that censoring is from above).

Whatever the value of $s$, we see that discrimination between $H_1$ and $H_2$ will be based on the statistic

$$T = \prod_{j=1}^{m} \{1-F(X_{1j},r_{1j})\}\{1-F(X_{2j},r_{2j})\}^{-1} = \prod_{j=1}^{m} (Z_{1j,r_{1j}}/Z_{2j,r_{2j}}) \tag{2}$$

where $Z_{ij,r} = 1 - F(X_{ij,r})$. Generally $H_1(H_2)$ is accepted if $T >(<) K$. If $r_{1j} = r_{2j}$ for all $j$ the likelihood ratio is just $T^s$ and it is natural to take $K=1$ and use the rule

$$\begin{aligned} &\text{"Accept } H_1 \text{ if } T>1. \\ &\text{Accept } H_2 \text{ if } T<1." \end{aligned} \tag{3}$$

(If $T=1$, no decision is reached.)

If $H_1$ is valid $Z_{1j,r_{1j}}$ has a standard beta distribution with parameters $(s+1),r_{1j}$ and $Z_{2j,r_{2j}}$, independent of $Z_{1j,r_{1j}}$, has a standard beta distribution with parameters $1,r_{2j}$ .

Each of the ratios $V_j = Z_{1j,r_{1j}}/Z_{2j,r_{2j}}$ is distributed as the ratio of independent standard beta variables with parameters $(s+1),r_{1j}$ and $1,r_{2j}$ respectively. The $V_j$'s are mutually independent and we have

$$\Pr[\text{correct decision}|H_1] = \Pr[\prod_{j=1}^{m} V_j > K].$$

When $m=1$ and $r_1 = r_2 = r$ we have

$$\Pr[\text{correct decision}|H_1] = \Pr[V_1 > 1]$$

$$= \frac{(s+r)!\,r!}{s!\{(r-1)!\}^2} \int_0^1 \int_0^{z_1} z_1^s (1-z_1)^{r-1} (1-z_2)^{r-1} dz_2 dz_1$$

$$= \frac{(s+r)!\,r!}{s!\{(r-1)!\}^2} \int_0^1 z_1^s (1-z_1)^{r-1} \cdot r^{-1}\{1-(1-z_1)^r\} dz_1$$

$$= \frac{(s+r)!\,r!}{s!\{(r-1)!\}^2} \cdot \frac{1}{r}[B(s+1,r) - B(s+1,2r)]$$

$$= \frac{(s+r)!\,r!}{s!\{(r-1)!\}^2} \cdot \frac{1}{r}\left[ \frac{s!(r-1)!}{(s+r)!} - \frac{s!(2r-1)!}{(s+2r)!} \right]$$

$$= 1 - \frac{(s+r)!(2r-1)!}{(r-1)!(s+2r)!} . \tag{4}$$

This is, of course, also the probability of correct decision when $H_2$ is valid. Some numerical values are given in Table 1.

TABLE 1

Probability of Correct Decision,
Based on Single Pair of Samples, Each of Size r

| s/r | 5 | 10 | 15 | 20 | 25 | 50 | ∞ |
|---|---|---|---|---|---|---|---|
| 1 | 0.727 | 0.738 | 0.742 | 0.744 | 0.745 | 0.748 | 0.750 |
| 2 | 0.841 | 0.857 | 0.863 | 0.866 | 0.868 | 0.871 | 0.875 |
| 3 | 0.902 | 0.919 | 0.925 | 0.928 | 0.930 | 0.934 | 0.9375 |
| 4 | 0.937 | 0.953 | 0.958 | 0.961 | 0.962 | 0.966 | 0.969 |
| 5 | 0.958 | 0.972 | 0.976 | 0.978 | 0.980 | 0.982 | 0.984 |

The probabilities in the above tables are, in fact, simply values of

$$\Pr[X_{1r} < X_{2r}|H_1] \tag{4'}$$

since $(V_1 > 1)$ is equivalent to $(X_{1r} < X_{2r})$.

For sequences $A_1, A_2$ of $m$ samples, when $m$ is large enough, we may use a normal approximation to the distribution of

$$T' = \sum_{j=1}^{m} \ell n \, V_j \ .$$

Using the methods described in Section 2.2 of [1] we obtain

$$\kappa_t(T'|H_1) = \sum_{j=1}^{m} [\{\psi^{(t-1)}(s+1) - \psi^{(t-1)}(r_{1j}+s+1)\}$$

$$+ (-1)^t \{\psi^{(t-1)}(1) - \psi^{(t-1)}(r_{2j}+1)\}] \qquad (5)$$

$$= (-1)^t (t-2)! \sum_{j=1}^{m} [\sum_{h=1}^{r_{1j}} (h+s)^{-(t-1)} + (-1)^t \sum_{h=1}^{r_{2j}} h^{-(t-1)}].$$

If $r_{1j} = r_{2j} = r$ for all $j = 1,2,\ldots m$, then

$$\kappa_t(T'|H_1) = m \cdot (t-2)! [\sum_{h=1}^{r} h^{-(t-1)} + (-1)^t \sum_{h=1}^{r} (h+s)^{-(t-1)}]. \qquad (6)$$

## 2.2 Population Distribution Not Known

Using the method described in section 3 of [1] we suppose that among $r_1$, $r_2$ observed values from samples in sequences $A_1$, $A_2$ respectively (with a common distribution) the $G_2$ greatest are from $A_2$ . Under $H_1$ , there were originally $(r_1 + s)$ values in a complete random sample $\overline{A}_1$ from which the greatest $s$ have been removed to form $A_1$ , while the $r_2$ values from $A_2$ represent a complete random sample. The number of order-ings of the original $(r_1 + r_2 + s)$ values which would produce an ordering with the $G_2$ greatest values from $A_2$ , after censoring in the way described is

$$\binom{G_2+s}{s}$$

and the corresponding likelihood is

$$L_1'/L_2' = \{(G_2+1)^{[s]}/(G_1+1)^{[s]}\} \cdot \{(r_1+1)^{[s]}/(r_2+1)^{[s]}\}. \tag{7}$$

If $r_1 = r_2 = r$ then the likelihood ratio is just

$$(G_2+1)^{[s]}/(G_1+1)^{[s]}. \tag{8}$$

Since one of $G_1$ and $G_2$ must be zero (and the other positive) we see that the likelihood ratio approach leads to the decision rule

$$\text{"Accept } H_i \text{ if } G_i > 0 \quad (i = 1,2)." \tag{9}$$

Since $G_1 > 0$ is equivalent to $X_{1r} > X_{2r}$ this rule is *equivalent to the rule (3) obtained from the approach based on knowledge of the population distribution.*

The probabilities of correct decision given in Table 1 thus apply, and we see that in this case we appear to lose nothing by not knowing the population distribution.

The situation is different when we have more than one sample in each sequence ($m>1$). If we suppose $r_{ij} = r$ for all $i$ and all $j$, we are led to the rule

$$\text{"Accept } H_1 \text{ if } \prod_{j=1}^{m} \left( \frac{(G_{2j}+1)^{[s]}}{(G_{1j}+1)^{[s]}} \right) > 1.$$

$$\text{Accept } H_2 \text{ if the ratio is less than 1."} \tag{10}$$

(If the ratio equals 1, no decision is reached.) The pairs $(G_{1j}, G_{2j})$ for $j = 1,\ldots m$ are mutually independent, and the joint distribution of $G_{1j}$ and $G_{2j}$ if $H_1$ is valid is

$$\left.\begin{array}{l} \Pr[(G_{1j}=g_1) \cap (G_{2j}=0) | H_1] = \binom{2r-g_1-1}{r-1} \Big/ \binom{2r+s}{r} \\[2ex] \Pr[(G_{1j}=0) \cap (G_{2j}=g_2) | H_1] = \binom{g_2+s}{s}\binom{2r-g_2-1}{r-1} \Big/ \binom{2r+s}{s} \end{array}\right\} \quad (11)$$

$(g_1, g_2 = 1, 2, \ldots, r)$.

The possible values of the ratio

$$W_j = (G_{2j} + 1)^{[s]} / (G_{1j} + 1)^{[s]}$$

are, in increasing order, $s!\{(g+1)^{[s]}\}^{-1}$ for $g = r, r-1, \ldots 1$ and $(s!)^{-1}(g+1)^{[s]}$ for $g = 1, 2, \ldots r$.

For small values of $m$ and $r$ it is possible to evaluate

$$\Pr[\text{correct decision}] = \Pr[\text{correct decision}|H_1] = \Pr[\prod_{j=1}^{m} W_j > 1 | H_1]. \quad (12)$$

For example if $m=2$, we have

$$\Pr[\text{correct decision}] = \Pr[\text{correct decision on each of } (A_{11}, A_{21})$$
$$\text{and } (A_{12}, A_{22}) \text{ separately}]$$
$$+ \Pr[\text{one correct and one incorrect decision}]$$
$$\times \Pr[W_1 W_2 > 1 | W_1 > 1; W_2 < 1]$$

$$= \left\{1 - \frac{(s+r)!(2r-1)!}{(s+2r)!(r-1)!}\right\}^2 + 2\binom{2r+s}{r}^{-2} \sum_{1 \leq g_1 < g_2 \leq r} \binom{g_2+s}{s}\binom{2r-g_2-1}{r-1}\binom{2r-g_1-1}{r-1} \quad (\text{using } (4) \text{ and } (11)). \quad (13)$$

The double summation is conveniently evaluated as

$$\sum_{g_2=2}^{r} \binom{g_2+s}{s}\binom{2r-g_2-1}{r-1} \sum_{g_1=1}^{g_2-1} \binom{2r-g_1-1}{r-1}. \quad (14)$$

With $m>1$ there is the possibility of no decision being reached. For $m=2$, the probability that no decision is reached is

$$2\binom{2r+s}{r}^{-2} \sum_{g=1}^{r} \binom{g+s}{s}\binom{2r-g-1}{r-1}^2 .$$  (15)

TABLE 2

Probabilities of Correct Decision and No Decision when m=2

| s/r | 5 | | 10 | | 15 | |
| --- | --- | --- | --- | --- | --- | --- |
| | Correct | No | Correct | No | Correct | No |
| 1 | 0.692 | 0.068 | 0.717 | 0.058 | 0.726 | 0.055 |
| 2 | 0.842 | 0.039 | 0.870 | 0.030 | 0.880 | 0.027 |
| 3 | 0.915 | 0.023 | 0.940 | 0.015 | 0.948 | 0.012 |
| 4 | 0.953 | 0.013 | 0.972 | 0.007 | 0.977 | 0.005 |
| 5 | 0.972 | 0.008 | 0.986 | 0.004 | 0.990 | 0.003 |

3.  Ceneral Censoring of Extremes

If we do not know the numbers  $s'$, $s''$  of least and greatest observations removed from one of two samples (or from each sample in one of two sequences) — though we do know that such removal *has* taken place from just one of them, we cannot use the direct likelihood ratio approach. Results obtained in [2] suggest that we might use criteria based on quantities $D_{ij} = G_{ij} + L_{ij}$ where $G_{ij}(L_{ij})$ is equal to the number of values among the sample $A_{ij}$ which are greater(less) than any values in $A_{3-i,j}$ .

If we simply have one pair of samples  $(m=1)$  of equal size  $r$, then we would:

"Accept $H_1$ if  $G_1 + L_1 < G_2 + L_2$

Accept $H_2$ if  $G_1 + L_1 > G_2 + L_2$

Reach no decision if  $G_1 + L_1 = G_2 + L_2$ ."

This is the same test as that obtained in [1] (see (30'), page 13) for the case when the censoring is known to be symmetrical.

If m is greater than 1, the value of $D_{ij}$ might be combined by either multiplication or summation, leading to the rules:

or

$$\text{"Accept } H_1(H_2) \text{ if } \prod_{j=1}^{m} (D_{ij} + 1) <(>) \prod_{j=1}^{m} (D_{2j} + 1)" \qquad (16)$$

$$\text{Accept } H_1(H_2) \text{ if } \prod_{j=1}^{m} D_{1j} <(>) \sum_{j=1}^{m} D_{2j} \text{ ."} \qquad (17)$$

In both (16) and (17), no decision is reached when there is equality. Neither of these rules is the same as that suggested for symmetrical censoring in [1] (foot of page 12).

An approximate assessment of the properties of procedure (17) can be based on normal approximation to the distribution of

$$T_m = \sum_{j=1}^{m} (D_{1j} - D_{2j}).$$

The first and second moments of $(D_{1j} - D_{2j})$ under $H_1$ are derived in the Appendix (equations (A12) and (A27) ). Numerical values for the expected value and variance of $(D_{1j} - D_{2j})$ under $H_1$ are given in Table 3. The variance under $H_2$ is the same as under $H_1$ , while the expected value has its sign reversed (positive under $H_2$ , negative under $H_1$ ).

## TABLE 3
### Moments of $(D_1 - D_2)$

| r | s' | s" | $E[D_1 - D_2 | H_1]$ | $\text{var}(D_1 - D_2 | H_1)$ | $\text{s.d.}(D_1 - D_2 | H_1)$ | m* | m** |
|---|---|---|---|---|---|---|---|
| 15 | 2 | 2 | -3.852 | 8.998 | 3.000 | 4 | 6 |
|  | 3 | 1 | -3.709 | 9.085 | 3.014 | 4 | 7 |
|  | 4 | 0 | -3.234 | 8.912 | 2.985 | 5 | 9 |
|  | 2 | 1 | -3.078 | 9.323 | 3.053 | 6 | 10 |
|  | 3 | 0 | -2.682 | 9.010 | 3.002 | 7 | 12 |
|  | 1 | 1 | -2.271 | 9.568 | 3.093 | 11 | 18 |
|  | 2 | 0 | -2.014 | 9.177 | 3.029 | 13 | 22 |
|  | 1 | 0 | -1.163 | 9.340 | 3.056 | 38 | 66 |
|  | 0 | 0 | 0 | 9.015 | 3.002 | - | - |
| 10 | 2 | 1 | -2.647 | 7.589 | 2.755 | 6 | 11 |
|  | 3 | 0 | -2.243 | 7.208 | 2.685 | 8 | 14 |
|  | 1 | 1 | -1.986 | 7.958 | 2.821 | 11 | 20 |
|  | 2 | 0 | -1.726 | 7.546 | 2.745 | 14 | 25 |
|  | 1 | 0 | -1.024 | 7.899 | 2.810 | 41 | 72 |
|  | 0 | 0 | 0 | 7.818 | 2.796 | - | - |
| 5 | 1 | 1 | -1.336 | 4.750 | 2.179 | 15 | 26 |
|  | 2 | 0 | -1.059 | 4.339 | 2.083 | 21 | 37 |
|  | 1 | 0 | -0.907 | 4.229 | 2.057 | 28 | 50 |
|  | 0 | 0 | 0 | 5.063 | 2.250 | - | - |

For $m$ modestly large — say in $\geq 4$ — the distribution of

$$T_m = \sum_{j=1}^{m} (D_{1j} - D_{2j})$$

will be approximately normal, and the probability of correct decision, given either $H_1$ or $H_2$ , will be approximately

$$\Phi\left(\sqrt{m} \left| E[D_1 - D_2 | H_1] \right| / \{\text{var}(D_1 - D_2 | H_1)\}^{\frac{1}{2}}\right). \tag{18}$$

In the last two columns of Table 3 **m*** gives the minimum integer value of $m$ for which the approximate probability of correct decision is at least 0.99, ie

$$\frac{m^* \{E[D_1 - D_2 | H_1]\}^2}{\text{var}(D_1 - D_2 | H_1)} \geq 2.326 \tag{19}$$

and $m^{**}$ gives the minimum value for which the approximate probability of correct decision is at least $0.999$.

As $r \to \infty$, ( $s'$ and $s''$ remaining constant) the limiting values of expected value and variance are

$$\lim_{r \to \infty} E[D_1 - D_2 | H_1] = -(s' + s' + 2) + 2^{-s'} + 2^{-s''} \qquad (20)$$

and

$$\lim_{r \to \infty} var(D_1 - D_2 | H_1) = 2(s' + s'' + 2) + (2s' + s)2^{-s'} + (2s'' + s)s^{-s''} - 4^{-s'} - 4^{-s''}. \qquad (21)$$

(We use

$$\lim_{r \to \infty} \{ \binom{2r+a}{r+b} / \binom{2r+a+c}{r+d} \} = 2^{-c} \lim_{r \to \infty} \{ \binom{2r+a}{r+b} / \binom{2r+a}{r+c} \} = 1,$$

$$\lim_{r \to \infty} \{ \binom{r+a}{r+b} / \binom{2r+c}{r+d} \} = 0. )$$

Table 4 gives some numerical values of these limits and corresponding values of $m^*$ and $m^{**}$. From Table 4 it appears that the limiting values are not closely approached unless $r$ is about 50 or more.

### TABLE 4

Limiting Expected Value and Variance as $r \to \infty$ ($s'$, $s''$ constant)

| $s'$ | $s''$ | $\lim_{r \to \infty} E[D_1 - D_2 \vert H_1]$ | $\lim_{r \to \infty} var(D_1 - D_2 \vert H_1)$ | $m^*$ | $m^{**}$ |
|---|---|---|---|---|---|
| 2 | 2 | -5.5 | 16.375 | 3 | 6 |
| 3 | 1 | -5.375 | 16.609375 | 4 | 6 |
| 4 | 0 | -4.984375 | 16.8046875 | 4 | 7 |
| 2 | 1 | -4.25 | 15.484375 | 5 | 9 |
| 3 | 0 | -3.875 | 15.359375 | 6 | 10 |
| 1 | 1 | -3.0 | 14.50 | 9 | 16 |
| 2 | 0 | -2.75 | 14.1875 | 11 | 18 |
| 1 | 0 | -1.5 | 13.25 | 32 | 57 |
| 0 | 0 | 0.0 | 12.0 | - | - |

The decision rule (16) is just as simple to apply as (17), but the distribution theory associated with it is rather more complex.

We can examine the relationship between (16) and (17) in a simple case ($r=2$, $s'=1$, $s''=0$) by direct calculation of probabilities. In this case there are $\binom{5}{2} = 10$ possible configurations of the combined sample $\bar{A}_1 \cap A_2$. In seven of these $D_1 = D_2$; in two $D_1 = 0$ and $D_2 = 2$, and in one $D_1 = 2$ and $D_2 = 0$.

Now (16) and (17) each lead to acceptance of $H_1$ (which is valid since $s'>0$) when there are more pairs of samples with one of the two ($D_1 = 0$, $D_2 = 2$) configurations than there are with the single ($D_1 = 2$, $D_2 = 0$) configuration. There will be no decision when the number of pairs of samples is the same for each of the two types of configuration.

The probability of correct decision is

$$\sum_{j=1}^{m} \binom{m}{j} \left(\frac{7}{10}\right)^{m-j} \left(\frac{3}{10}\right)^{j} \sum_{h=[\frac{1}{2}j]+1}^{j} \binom{j}{h} \left(\frac{2}{3}\right)^{h} \left(\frac{1}{3}\right)^{j-h}$$

$$= (0.7)^{m} \sum_{j=1}^{m} \binom{m}{j} 7^{-j} \sum_{h=[\frac{1}{2}j]+1}^{j} \binom{j}{h} 2^{h} \quad (22)$$

where $[\frac{1}{2}j]$ denotes the integer part of $\frac{1}{2}j$.

The probability of no decision is

$$\sum_{j=0}^{[\frac{1}{2}m]} \binom{m}{2j} \left(\frac{7}{10}\right)^{m-2j} \left(\frac{3}{10}\right)^{2j} \binom{2j}{j} \left(\frac{2}{3}\right)^{j} \left(\frac{1}{3}\right)^{j}$$

$$= (0.7)^{m}(m!) \sum_{j=0}^{[\frac{1}{2}m]} \{(m-2j)!\}^{-1} (j!)^{-2} \left(\frac{2}{49}\right)^{j} . \quad (23)$$

(Of course, we take $\binom{0}{0} = 1$.) Alternatively, from the point of view of procedure (17), we need the distribution of $T_m = \sum_{j=1}^{m} (D_{1j} - D_{2j})$. The

distribution of $(D_{1j}-D_{2j})$ is

$$Pr[D_{1j}-D_{2j} = d|H_1] = \begin{cases} \dfrac{1}{5} & \text{for } d=-2 \\[2mm] \dfrac{7}{10} & \text{for } d=0 \\[2mm] \dfrac{1}{10} & \text{for } d=2 \end{cases}$$

and the probability generating function is

$$\frac{1}{10}(2t^{-2} + 7 + t^2) = \frac{1}{10} t^{-2}(2 + 7t^2 + t^4).$$

The probability generating function of $T_m$ is $10^{-m}t^{-2m}(2 + 7t^2 + t^4)^m$. Hence the probability of correct decision when $H_1$ is valid, using (17), is

$$Pr[T_m < 0|H_1] = 10^{-m} \times \text{(sums of coefficients in terms of } t^\alpha$$
$$\text{with } \alpha < 2m \text{ in } (2+7t^2+t^4)^m ). \tag{24}$$

The probability of no decision, when using (17) is

$$Pr[T_m = 0|H_1] = 10^{-m}\text{(coefficient of } t^{2m} \text{ in } (2+7t^2+t^4)^m ). \tag{25}$$

Table 5 contains a few numerical values. The values given by (22) and (24), and by (23) and (25) are of course identical.

In general, procedures (16) and (17) will not be identical, but this example is of interest in showing that they can be.

TABLE 5

Decision Probabilities when $H_1$ (or $H_2$) is Valid
Under Either Rules (16) or (17) when $r=2$, $s'=1$, $s''=0$

| m | Correct | No |
|---|---------|------|
| 2 | 0.32 | 0.53 |
| 4 | 0.45 | 0.36 |
| 6 | 0.53 | 0.28 |

## APPENDIX

### Moments of $(D_1 - D_2)$

(i) We will use the convention $\binom{a}{b} = 0$ if $a < b$.

(ii) We will use the relationships $(a < b)$

$$\sum_{j=a}^{b} \binom{j}{r} = \binom{b+1}{r+1} - \binom{a}{r+1} \tag{A1}$$

and

$$\sum_{j=0}^{r} \binom{a+j}{a}\binom{b+r-j}{b} = \binom{a+b+1+r}{a+b+1} = \binom{a+b+1+r}{r}. \tag{A2}$$

For convenience we will omit the subscript denoting order in the series and consider

$$D_1 - D_2 = L_1 + G_1 - L_2 - G_2$$

where $L_i(G_i)$ is the number of values in $A_i$ less(greater) than any values in $A_{3-i}$.

We introduce indicator variables

$$_L X_{ih} = \begin{cases} 1 & \text{if the } h\text{-th order statistic of } A_i \text{ is less than any value in } A_{3-i} \\ 0 & \text{otherwise} \end{cases}$$

and $_G X_{ih}$, defined as $_L X_{ih}$ but with 'less' replaced by 'greater'.
Then

$$L_i = \sum_{h=1}^{r} {}_L X_{ih} \; ; \quad G_i = \sum_{h=1}^{r} {}_G X_{ih} \tag{A3}$$

and

$$E[{}_L X_{ih}^{\alpha}] = Pr[{}_L X_{ih} = 1]; \quad E[{}_G X_{ih}^{\alpha}] = Pr[{}_G X_{ih} = 1] \qquad (\alpha = 1,2)$$

$$E[{}_L X_{ih} \, {}_L X_{ik}] = E[{}_L X_{ik}] \quad \text{for } k \geq h \tag{4.1}$$

$$E[{}_G X_{ih} \, {}_G X_{ik}] = E[{}_G X_{ih}] \quad \text{for } k \geq h. \tag{4.2}$$

Under $H_1$ , the total number of ways of arranging the original $(r+s'+s'')$ values in $\overline{A}_1$ and the $r$ values of $A_2$ in ascending order of magnitude is $\binom{2r+s'+s''}{r}$. Consequently

$$E[_LX_{ih}|H_1] = \frac{\text{Number of arrangements for which } _LX_{ih}=1}{\binom{2r+s'+s''}{r}}$$

and similarly for expected values of other quantities. It is convenient to introduce the symbols

$$E^*[\cdot|H_1] = \binom{2r+s'+s''}{r}E[\cdot|H_1] \tag{A5}$$

in our calculations.

We will use pictorial representation to indicate the derivation of the various combinatorial formulae.

$$E^*[_LX_{1h}|H_1] = \binom{2r+s''-h}{r} \quad \begin{array}{c|c|c} \overline{A}_1 & s'+h & r+s''-h \\ \hline A_2 & 0 & r \end{array} \tag{A6}$$

$$E^*[_LX_{2h}|H_1] = \sum_{u=0}^{s'} \binom{u+h-1}{h-1}\binom{2r+s'+s''-u-h}{r-h} \quad \begin{array}{c|c|c} \overline{A}_1 & u\leq s & r+s'+s''-u \\ \hline A_2 & h-1\times & r-h \end{array} \tag{A7}$$

From (A3) and (A6)

$$E^*[L_1|H_1] = \sum_{h=1}^{r} \binom{2r+s''-h}{r} = \binom{2r+s''}{r+1} - \binom{r+s''}{r+1} \tag{A8}$$
$$\text{(using (A1) )}.$$

Similarly

$$E^*[G_1|H_1] = \binom{2r+s'}{r+1} - \binom{r+s'}{r+1}. \tag{A9}$$

From (A3) and (A7)

$$E^*[L_2|H_1] = \sum_{u=0}^{s'} \sum_{h=1}^{r} \binom{u+h-1}{h-1}\binom{2r+s'+s''-u-h}{r-h}$$

$$= \sum_{u=0}^{s'} \binom{2r+s'+s''}{r-1} \qquad \text{(using (A2))}$$

$$= (s'+1)\binom{2r+s'+s''}{r-1}. \tag{A10}$$

Similarly

$$E^*[G_2|H_1] = (s''+1)\binom{2r+s'+s''}{r-1}. \tag{A11}$$

From (A8) - (A11)

$$E[D_1-D_2|H_1] = \frac{\binom{2r+s'}{r+1}+\binom{2r+s''}{r+1}-\binom{r+s'}{r+1}-\binom{r+s''}{r+1}-(s'+s''+2)\binom{2r+s'+s''}{r-1}}{\binom{2r+s'+s''}{r}}. \tag{A12}$$

We now turn to the evaluation of

$$E[(D_1-D_2)^2|H_1] = E[L_1^2|H_1] + E[G_1^2|H_1] + E[L_2^2|H_1] + E[G_2^2|H_1] + 2E[L_1G_1|H_1]$$

$$+ 2E[L_2G_2|H_1] - 2E[L_1G_2|H_1] - 2E[L_2G_1|H_1]. \tag{A13}$$

(Note that $L_1L_2$ are $G_1G_2$ are always zero.)

In view of (A4.1)

$$E^*[L_i^2|H_1] = E[\sum_{h=1}^{r} {}_LX_{ih}^2 + 2\sum_{h}^{r-1}\sum_{k}^{r} {}_LX_{ih}\,{}_LX_{ik}|H_1]$$

$$= E^*[L_i + 2\sum_{k=1}^{r}(k-1)\,{}_LX_{ih}|H_1]. \tag{A14}$$

Now

$$E^*[\sum_{k=1}^{r}(k-1)\,{}_LX_{1k}|H_1] = \sum_{k=1}^{r}(k-1)\binom{2r+s''-k}{r}$$

$$= \sum_{j=0}^{r-2}\sum_{\mu=0}^{j}\binom{r+s''+\mu}{r}$$

$$= \sum_{j=0}^{r-2}\{\binom{r+s''+j+1}{r+1}-\binom{r+s''}{r+1}\}$$

$$= \binom{2r+s''}{r+2} - \binom{r+s''+1}{r+2} - (r-1)\binom{r+s''}{r+1}$$

$$= \binom{2r+s''}{r+2} - \binom{r+s''}{r+2} - r\binom{r+s''}{r+1}$$

and so

$$E^*[L_1^2|H_1] = \binom{2r+s''}{r+1} - \binom{r+s''}{r+1} + 2\left\{\binom{2r+s''}{r+2} - \binom{r+s''}{r+2} - r\binom{r+s''}{r+1}\right\}$$

$$= \binom{2r+s''}{r+1} + 2\binom{2r+s'}{r+2} - (2r+1)\binom{r+s'}{r+1} - 2\binom{r+s'}{r+2}. \tag{A15}$$

Similarly

$$E[G_1^2|H_1] = \binom{2r+s'}{r+1} + 2\binom{2r+s'}{r+2} - (2r+1)\binom{r+s'}{r+1} - 2\binom{r+s'}{r+2}. \tag{A16}$$

Also

$$E^*\left[\sum_{k=1}^{r} (k-1)\,_L X_{2k}|H_1\right] = \sum_{k=1}^{r} (k-1) \sum_{u=0}^{s'} \binom{u+k-1}{k-1}\binom{2r+s'+s''-u-k}{r-k}$$

$$= \sum_{u=0}^{s'} (u+1) \sum_{k=2}^{r} \binom{u+k-1}{u+1}\binom{2r+s'+s''-u-k}{r-h}$$

$$= \binom{2r+s'+s''}{r+s'+s''+2} \sum_{u=0}^{s'} (u+1) \qquad \text{(using (A2))}$$

$$= \tfrac{1}{2}(s'+1)(s'+2)\binom{2r+s'+s''}{r-2} \tag{A17}$$

and so

$$E^*[L_2^2|H_1] = (s'+1)\binom{2r+s'+s''}{r-1} + (s'+1)(s'+2)\binom{2r+s'+s''}{r-2}. \tag{A18}$$

Similarly

$$E^*[G_2^2|H_1] = (s''+1)\binom{2r+s'+s''}{r-1} + (s''+1)(s''+2)\binom{2r+s'+s''}{r-2}. \tag{A19}$$

To evaluate $E^*[L_1 G_1|H_1]$ we need (with $k>h$)

$$E^*[_LX_{ih}\ _GX_{ik}|H_1] = \binom{r+k-h-1}{r}. \tag{A20}$$

| $\overline{A}_1$ | $s'+h$ | $k-h-1$ | $s''+r-k+1$ |
|---|---|---|---|
| $A_2$ | 0 | $r$ | 0 |

From this

$$E^*[L_1G_1|H_1] = \sum_{h=1}^{r-1}\ \sum_{k=h+1}^{r}\ \binom{r+k-h-1}{r}$$

$$= \sum_{h=1}^{r-1}\ \binom{2r-h}{r+1} = \binom{2r}{r+2} \tag{A21}$$

(using (A1) twice).

Also, with $k>h$ again

$$E^*[_LX_{2h}\ _GX_{2k}|H_1] = \sum_{u=0}^{s'}\ \sum_{v=0}^{s''}\ \binom{u+h-1}{h-1}\binom{r+s'+s''-u-v+k-h-1}{k-h-1}\binom{v+r-k}{r-k}. \tag{A22}$$

| $\overline{A}_1$ | $u\le s'$ | $r+s'+s''-u-v$ | $v\le s''$ |
|---|---|---|---|
| $A_2$ | $h-1$ $\times$ | $k-h-1$ | $\times r-k$ |

Hence

$$E^*[L_2G_2|H_1] = \sum_{u=0}^{s'}\ \sum_{v=0}^{s''}\ \sum_{k=2}^{r}\ \binom{v+r-k}{r-k}\ \sum_{h=1}^{k=1}\ \binom{u+h-1}{u}\binom{r+s'+s''-u-v+k-h-1}{r+s'+s''-u-v}$$

$$= \sum_{u=0}^{s'}\ \sum_{v=0}^{s''}\ \sum_{k=2}^{r}\ \binom{v+r-k}{v}\binom{r+s'+s''-v+k-1}{k-2}$$

(using (A2))

$$= \sum_{u=0}^{s'}\ \sum_{v=0}^{s''}\ \binom{2r+s'+s''}{r-2}$$

(using (A2) again)

$$= (s'+1)(s''+1)\binom{2r+s'+s''}{r-2}. \tag{A23}$$

To calculate $E^*[L_1G_2|H_1]$ we need (for _any_ h,k = 1,...r)

$$E^*[_L X_{ih} \; _G X_{2k} | H_1] = \sum_{v=0}^{s''} \binom{v+r-k}{v} \binom{r+s''-v-h+k-1}{r+s''-v-h}$$  (A24)

| | $\bar{A}_1$ | $s'+h$ | $r+s''-v-h$ | $v \le s''$ |
|---|---|---|---|---|
| | $A_2$ | $0$ | $k-1$ | $r-k$ |

whence

$$E^*[_L L_1 G_2 | H_1] = \sum_{v=0}^{s''} \sum_{h=1}^{r} \sum_{k=1}^{r} \binom{v+r-k}{v} \binom{r+s''-v-h+k-1}{r+s''-v-h}$$

$$= \sum_{v=0}^{s''} \sum_{h=1}^{r} \binom{2r+s''-h}{r-1} \qquad \text{(using (A2))}$$

$$= \sum_{v=0}^{s''} \binom{2r+s''}{r} - \binom{r+s''}{r} \qquad \text{(using (A1))}$$

$$= (s''+1)\{\binom{2r+s''}{r} - \binom{r+s''}{r}\}.$$  (A25)

Similarly

$$E^*[_G G_1 L_2 | H_1] = (s'+1)\{\binom{2r+s'}{r} - \binom{r+s'}{r}\}.$$  (A26)

Combining (A13, 15, 16, 13, 19, 21, 23, 25, 26) we obtain after some rearrangement

$$E[(D_1-D_2)^2 | H_1] = \binom{2r+s'+s''}{r}^{-1} [\sum{}' \{2\binom{2r+s}{r+2} + \binom{2r+s}{r+1} - 2(s+1)\binom{2r+s}{r}$$

$$- 2\binom{r+s}{r+2} - (2r+1)\binom{r+s}{r+1} + 2(s+1)\binom{r+s}{r}\}$$

$$+ (s'+s''+2)\binom{2r+s'+s''}{r-1} + (s'+s''+2)(s'+s''+3)\binom{2r+s'+s''}{r-2}$$

$$+ 2\binom{2r}{r-2}]$$  (A27)

where $\sum{}'$ denotes summation over $s=s'$, $s=s''$.

As a check we note that for $r=1$, when $L_1 + G_1 = L_2 + G_2 = 1$ and as $D_1 - D_2 = 0$ identically, (A12) gives $E[D_1-D_2|H_1] = 0$ and (A27) gives $\text{var}(D_1-D_2|H_1) = 0$.

Of course, if in a sequence of sample pairs each of size 1 we had a preponderance of pairs with $L_1 = 1 = G_2$ we would suspect that *one* of the two samples had been censored unsymmetrically, but we would not be able to decide *which* one.

If $s' = s'' = 0$ then, as we would expect, from (A12)

$$E[D_1-D_2|H_1] = 0 \qquad \text{(for any r)}$$

and from (A27)

$$\text{var}(D_1-D_2|H_1) = 4\binom{2r}{r}^{-1} + \frac{4(3r^2-4r-2)}{(r+1)(r+2)} . \qquad \text{(A28)}$$

## Distribution of $(L_1, G_1, L_2, G_2)$

The joint distribution of $L_1$, $G_1$, $L_2$, $G_2$ when $H_1$ is valid is given by the following table.

| $\ell_1$ | $g_1$ | $\ell_2$ | $g_2$ | $\binom{2r+s'+s''}{r}\Pr[\bigcap\limits_{i=1}^{2}(L_i=\ell_i)\bigcap\limits_{i=1}^{2}(G_i=g_i)]$ |
|---|---|---|---|---|
| $\le 0$ | $>0$ | 0 | 0 | $\binom{2r-2-\ell_1-g_1}{r-2}$ |
| $>0$ | 0 | 0 | $>0$ | $\binom{2r-2-\ell_1-g_2}{r-\ell_1-1} \cdot \binom{s''+g_2}{s''}$ |
| 0 | $>0$ | $>0$ | 0 | $\binom{2r-2-\ell_2-g_1}{r-g_1-1} \cdot \binom{s'+\ell_2}{s'}$ |
| 0 | 0 | $>0$ | $>0$ | $\binom{2r-2-\ell_2-g_2}{r-2} \cdot \binom{s'+\ell_2}{s'} \cdot \binom{s''+g_2}{s''} .$ |

The joint distribution of $D_1 = L_1 + G_1$ and $D_2 = L_2 + G_2$ is given by the following table

| $d_1$ | $d_2$ | $\binom{2r+s'+s''}{r}\Pr[(D_1=d_1)\cap(D_2=d_2)]$ |
|-------|-------|---|
| $>0$ | $>0$ | $\binom{2r-2-d_1-d_2}{r-1-d_1}\left\{\binom{s'+d_2}{d_2}+\binom{s''+d_2}{d_2}\right\}$ |
| $>0$ | $0$ | $(r-1)\binom{2r-2-d_1}{r-2}$ |
| $0$ | $>0$ | $\binom{2r-2-d_2}{r-2}\sum_{j=1}^{d_2-1}\binom{s'+j}{s}\binom{s''+d_2-j}{s''}$ |

$$= \binom{2r-2-d_2}{r-2}\left\{\binom{s'+s''+1+d_2}{d_2}-\binom{s'+d_2}{d_2}-\binom{s''+d_2}{d_2}\right\}.$$

Calculation of the moments of $(D_1-D_2)$ from the joint distribution of $D_1$ and $D_2$ does not seem to be as convenient as calculation using indicator variables. This method is easily extended to the case when $r_1$ and $r_2$ , the sample sizes in $A_1$, $A_2$ respectively, are not necessarily equal. However, the appropriate test criteria would then be different, and so the results will not be given here.

Tukey [3] uses a test statistic (for differences between two samples) which in our notation, would be defined as: — if $L_iG_{3-i} \neq 0$, the value of the statistic is $(L_i + G_{3-i})$. (If $L_1G_2 = L_2G_1 = 0$, no value is assigned to the statistic.) The reader may find the derivation of the distribution of the statistic in [3] of interest.

# REFERENCES

[1] Johnson, N. L. (1976) "Completeness comparisons among sequences of samples." *Institute of Statistics, University of North Carolina, Mimeo Series No. 1056.* (To appear in *H. O. Hartley 65th Birthday Volume.*)

[2] Johnson, N. L. (1970) "A general purpose test of censoring of sample extreme values" (In *Essays in Probability and Statistics (S. N. Roy Memorial Volume)*) Chapel Hill: University of North Carolina Press (pp. 379-384).

[3] Tukey, J. W. (1959) "A quick, compact two-sample test to Duckworth's specifications" *Technometrics,* $\underline{1}$, 21-48.